Exploring Regression Structure with Graphics[1]

By

R. Dennis Cook[2] and Nate Wetzel

School of Statistics, University of Minnesota
Technical Report #595
October, 1993

[1] Penultimate version of an invited paper to appear with discussion in *TEST*.

[2] Research supported in part by grant DMS-9212413 from the National Science Foundation.

Exploring Regression Structure with Graphics[1]

BY

R. Dennis Cook[2] and Nate Wetzel

Technical Report 595

School of Statistics

University of Minnesota

October 15, 1993

Abstract

We investigate the extent to which it may be possible to carry out a regression analysis using graphics alone, an idea that we refer to as *graphical regression*. The limitations of this idea are explored. It is shown that graphical regression is theoretically possible with essentially no constraints on the conditional distribution of the response given the predictors, but with some conditions on marginal distribution of the predictors. Dimension reduction subspaces and added variable plots play a central role in the development. The possibility of useful methodology is explored through two examples.

---

[1]Penultimate version of an invited paper to appear with discussion in *TEST*.

[2]Research supported in part by grant DMS-9212413 from the National Science Foundation.

1. Introduction

Until about 10 years ago our personal views of statistical graphics encompassed little more than the displays that could be produced on a teletype or CRT terminal. The idea of interactively manipulating a plot or using motion on a computer screen to add another dimension did not have an operational meaning. The situation is drastically different today. There is now much commercial and public-domain software that allows access to many novel graphical techniques. Animation, brushing, grand tours, identification, linking, slicing, and spinning are some of the techniques that have greatly enhanced our ability to analyze data graphically.

We understand that much of what we have now can be traced back to the pioneering work on PRIM-9 (Fisherkeller, Friedman and Tukey, 1974) and to Peter Huber's visions for PRIM-ETH and PRIM-H. The collection edited by Cleveland and McGill (1988) contains a variety of useful papers on dynamic and interactive graphics from the late 1960's to the publication date. Cleveland (1987) gives a useful perspective on research in statistical graphics along with many references. A number of interesting remarks about the role of graphical methods in statistics, including a statement on the need for a graphical theory, are available in Cox (1978). For us, modern graphics became a concrete tool shortly after Luke Tierney began his work on *XLISP-STAT*, a programming environment that allows easy access to most of the modern techniques and, perhaps more importantly, allows the user to construct instances of new graphical ideas with relatively little difficulty (Tierney 1990).

In this paper we consider graphics for regression problems with the following structure: Let y_i denote the i-th observation on the univariate response variable y and let x_i denote the corresponding vector of observations on the $p\times 1$ vector of predictors x. We assume throughout that the data (y_i, x_i^T), i=1,...,n, are independent and identically distributed realizations on the random vector (y, x^T). Following an apparently standard convention, all cumulative distribution functions (cdf) will be denoted by F with the arguments indicating the random variables involved. The cdf for the conditional random variable y|x is denoted by F(y|x), for example.

A usual goal of regression analyses and the specific goal of this paper is to characterize how the distribution of y|x changes as the value of x ranges in the relevant sample space. There are of course many, many ways to pursue this goal, but we intentionally tie our hands and consider how progress might be made by using graphics alone. We will refer to this general idea as *graphical regression*. While we will try to make

this idea clear as the paper progresses, we will not be assuming specific functional representations for F(y|x). We will be estimating relevant characteristics of F(y|x), but not by using specific models or mathematical objective functions unless it appears that there is no other way to proceed. This rules out the possibility of using methods based on generalized additive models (Hastie and Tibshirani 1990) or projection pursuit indices (Huber 1985), for example. This should not be taken to imply any deficiencies in such methods, nor should it be taken to imply that graphical regression is not applicable in situations where such methods are appropriate.

Our investigation of graphical regression here is intended mainly as an academic exercise. By confining ourselves to a somewhat artificial framework we hope to learn more about potential roles for graphics in regression and with luck learn a little that has some immediate practical value. The motivation for this inquiry came from wondering about how far various graphical techniques could be pushed in an effort to understand regression problems. For example, consider a rotating three-dimensional scatterplot of y versus a pair of predictors, say x_j and x_k. With a little study the plot may provide useful information about how $F(y|x_j, x_k)$ varies with the values of the two predictors. But what information does this provide about the object of primary interest F(y|x) when there are p>2 predictors? Is a three-dimensional plot of y versus (x_j, x_k) a "best" graphical construction for inferring about characteristics of F(y|x)? Is it ever profitable or in some sense necessary to study all possible $\binom{p}{2}$ plots of this type?

A (q+1)-dimensional scatterplot will be denoted by {a, b} where the first argument, which will always be a scalar, is allocated to the vertical axis and the coordinates of the vector b are allocated to the "horizontal" axes in any convenient way. Scatterplots are conceptually viewed in this paper while rotating around the vertical axis. We assume that the reader is familiar with certain dynamic graphical techniques. Background on scatterplots is available in Cleveland (1984). Brushing, linking and allied operations are discussed in Becker and Cleveland (1987), Becker, Cleveland and Wilks (1987), and in Becker, Cleveland and Weil (1988). Background on viewing three-dimensional scatterplots via rotation is available in Becker, Cleveland and Weil (1988), Young, Kent, and Kuhfeld (1988), and Huber (1987). The use of dynamic graphics in the context of regression diagnostics is investigated by Cook and Weisberg (1989, 1990).

Following Dawid (1979) we use the notation $u \perp\!\!\!\perp v$ to indicate that the random variables u and v are independent. Similarly, $u \perp\!\!\!\perp v \mid z$ means that u and v are independent given any value for the random variable z. The subspace of R^q spanned by the columns of the q×t matrix A will be denoted by S(A), and dim(S) is the dimension of the subspace S.

In the next section we try to add some substance to the idea of graphical regression by considering problems in which there are just p=2 predictors. Many of the ideas in this section will find direct application in Section 3 which covers the many predictor case. The developments of Section 3 are based on selected results from Cook (1992, 1994) where the possibility of graphical regression is briefly introduced. Section 4 contains two short examples. In Section 5 we discuss various other issues to round out the discussion. Justifications are given mainly in terms of population calculations. Sample versions can always be constructed by substituting consistent estimates for the unknown quantities.

Finally, all of the graphical displays are based on one implementation of graphical regression ideas written in *XLISP-STAT*. We briefly describe characteristics of this implementation throughout the discussion.

2. Regression with Two Predictors

2.1 Introduction

We use the ethanol data as described in Cleveland and Hastie (1992) to introduce basic ideas in this section. The data are from an industrial experiment to study the exhaust from an experimental one-cylinder engine using ethanol as fuel. The response variable, NOx in *μg/joule*, is the concentration of nitrogen oxide plus nitrogen dioxide, normalized by the work of the engine. The two predictor variables are E, a measure of the richness of the air-fuel mixture, and the compression ratio, C. There are 88 observations on y = NOx and x^T=(E, C).

Since there are only two predictors we can view the entire data set in a three-dimensional scatterplot rotating about the vertical axis, {NOx, (C, E)}. One two-dimensional projection of this scatterplot is shown in Figure 1; the relevance of the highlighted points will be indicated later. Two things become immediately apparent as we rotate the plot: The distribution of NOx|(C, E) surely depends on the values of the predictors and E(NOx|C, E) is a nonlinear function of the values of C and E with a strong quadratic tendency. Clearly, there is ample evidence to contradict the possibility of a trivial regression in which NOx $\perp\!\!\!\perp$ (C, E).

Next imagine rotating the point cloud to the strongest, the most interesting or the "best" two-dimensional projection. For example, "best" might mean the projection with the smallest variation about a visually determined trend. This usually happens naturally. After viewing a rotating three-dimensional plot, people tend to stop at what they consider to be an interesting or striking two-dimensional view, often fine-tuning the view to obtain the "best" possible. In the present example, the nonlinear trend is clearly the most striking feature and

most will stop at the projection that seems to give the least visual variation about the nonlinear trend. The view that we selected is the one shown in the plot of Figure 1. The variable on the horizontal axis of a "best" two-dimensional projection is some linear combination of the predictors, say $b^T x$. For the projection of Figure 1, $b^T x = b_1 C + b_2 E$ where $b_1 = 0.01$ and $b_2 = 0.99$. The process thus far can be thought of as visually determining the best linear combination of the predictors with which to explain the variation in the response. A brief description might be visual fitting with a mental objective function.

At this point a crucial question arises: Is the projection in Figure 1 all that is necessary to characterize how the distribution of NOx|(C, E) varies with the predictors? Stated differently, is there information to contradict the conjecture

$$\text{NOx} \perp\!\!\!\perp (\text{C}, \text{E}) \mid (b_1 C + b_2 E) \tag{2.1}$$

If no such contradiction can be found then there may be no important loss of information when using the plot in Figure 1 as a substitute for the full rotating three-dimensional plot. Informally, we could then say that $\{y, (b_1 C + b_2 E)\}$ is a *sufficient* replacement for {NOx, (C, E)}. If evidence is found indicating that (2.1) is clearly false, then there are two possibilities: Either (2.1) is false for all linear combinations $b^T x$ of the predictors, or (2.1) is false for the particular linear combination at hand while it is true for some other linear combination. In the former case, no two-dimensional projection is sufficient and we must study the full three-dimensional plot to understand how the distribution of NOx|(C, E) varies with the value of (C, E). Further analysis may bring us closer to the desired plot in the latter case.

There are three possible outcomes of graphical regression with two predictors. These outcomes can be represented in terms of the single expression

$$y \perp\!\!\!\perp x \mid \eta^T x \tag{2.2}$$

for various values of the $2 \times q$ matrix η, $q \le 2$. First, if $\dim(S(\eta))=0$ and equation (2.2) holds then $y \perp\!\!\!\perp x$. Second, if $\dim(S(\eta))=1$ and (2.2) holds then the plot of y versus the linear combination of x determined by any basis for $S(\eta)$ is a sufficient two-dimensional projection of the full three-dimensional plot. Third, if $\dim(S(\eta))=2$ then (2.2) holds trivially. A basic task in graphically analyzing a three-dimensional plot is to determine the subspace $S_{y|x}$ of minimal dimension so that (2.2) holds for any basis η of the subspace. Following Cook (1994), we refer to such subspaces as *minimum dimension reduction subspaces* and let $d=\dim(S_{y|x})$. In order to keep track of these three situations, we say that the three-dimensional plot {y, x} exhibits *d-dimensional structure* for d=0,1,2.

Returning to the ethanol example, we know from Figure 1 that d>0 and we now must determine if d = 1 or 2: Is there is information in the data to contradict (2.2) when $S(\eta) = S((0.01, 0.99)^T)$?

2.2 Determining Dimension

It is often, but not always, easy to determine visually if d >0, as in the ethanol data. It may be easy on occasion to see from the rotating plot that d = 2. The choice between d =1 and d =2 is usually the most difficult, however, and in this section we discuss graphical methods to aid that choice.

2.2.1 A first method. One graphical method to determine d for the ethanol data can be based on a straightforward application of (2.1). Begin by forming a slice of highlighted points around a selected value for (b_1C+b_2E) as shown in Figure 1. If (2.1) is a good approximation of the data then the points in the slice should appear as an independent and identically distributed sample. To see if there is information to the contrary, rotate the point cloud with the points in the slice highlighted. They will appear as a rotating horizontal band of points under (2.1). Any clear systematic tendency may be taken as evidence to contradict (2.1). The plot in Figure 2 is a new two-dimensional projection of {NOx, (C,E)} with the slice points highlighted. There is surely a systematic pattern in the highlighted points and thus (2.1) does not hold for the particular linear combination $b^Tx = (b_1C+b_2E)$ that we selected. At this point we could rotate the plot in Figure 1 to a different projection in an attempt to remove systematic patterns like that in Figure 2, or we could concluded that d=2 and thus that $\{NOx, b_1C+b_2E\}$ is an *insufficient* plot.

If no systematic pattern had been detected in Figure 2, then it would be necessary to select a different slice and repeat the procedure. Failing to find a convincing systematic pattern in a series of slices that covers the range of (b_1C+b_2E), it may be reasonable to conclude that d=1.

There are two potential problems that we have noticed with this idea for direct application of (2.1). First, if the slices are not sufficiently fine, a remnant intraslice relationship may remain that can be confusing when the plot is rotated. Second, the procedure is awkward and it tends to be time consuming. This could be a problem because it may be necessary to apply it many times when there are many predictors, as described in Section 3. We tried the following approach in an effort to overcome these problems.

2.2.2 An attempt at improvement. The ideas in this and subsequent sections are described in terms of population quantities. Sample versions can be constructed as usual by substituting consistent estimates.

Let J_s denote the slice interval as illustrated in Figure 1. We assume that the slices are sufficiently narrow so that any intraslice dependence of $F(y|b^Tx)$ on b^Tx in the interval J_s can be described adequately by the location model $E(y|b^Tx) = \alpha_s + \beta_s(b^Tx)$, $b^Tx \in J_s$. Let $e_s = y - E(y|b^Tx)$ denote a typical population residual from an intraslice simple linear regression of y on b^Tx. Passing from the response to the residuals is intended to remove the remnant intraslice relationship mentioned above. Then it follows that

$$e_s \perp\!\!\!\perp b^Tx \mid (b^Tx \in J_s) \tag{2.3}$$

Next, choose a nonzero 2×1 vector b_s such that

$$b_s^T \operatorname{cov}(x \mid b^Tx \in J_s)b = 0 \tag{2.4}$$

and finally assume that

$$(e_s, b_s^Tx) \perp\!\!\!\perp b^Tx \mid (b^Tx \in J_s) \tag{2.5}$$

Then it follows from Dawid (1979) that

$$e_s \perp\!\!\!\perp (b_s^Tx, b^Tx) \mid (b^Tx \in J_s) \tag{2.6}$$

if and only if

$$e_s \perp\!\!\!\perp b_s^Tx \mid (b^Tx \in J_s) \tag{2.7}$$

The results in (2.6) and (2.7) are potentially useful for the following reasons. We need to assess condition (2.6) to determine if there is intraslice information to contradict d=1. But the straightforward method for doing this is laborious since it requires viewing a rotating plot $\{e_s, (b_s^Tx, b^Tx) \mid b^Tx \in J_s\}$ for each slice as described in Section 2.2.1. Under (2.3) - (2.5), however, an equivalent condition is given by (2.7). An assessment of this condition requires inspecting only the two-dimensional scatterplot $\{e_s, b_s^Tx \mid b^Tx \in J_s\}$.

A paradigm for deciding between d=1 and d=2 is as follows. Begin by rotating the three-dimensional plot {y, x} to the best two-dimensional projection $\{y, b^Tx\}$. Slice the two-dimensional plot $\{y, b^Tx\}$ around a value of b^Tx, construct the residuals e_s and the orthogonal direction b_s. And then send this information to the linked two-dimensional plot $\{e_s, b_s^Tx \mid b^Tx \in J_s\}$. The procedure can be automated easily so that brushing $\{y, b^Tx\}$ causes the quantities J_s, e_s and b_s to be recomputed in real time and the linked plot $\{e_s, b_s^Tx \mid b^Tx \in J_s\}$ to be updated.

In the interface shown in Figure 1, the procedure is initiated by clicking the "slice" button on the left of the plot. The two-dimensional plot $\{e_s, b_s^Tx \mid b^Tx \in J_s\}$ with sample

estimates of b, b_s, and e_s, and a slider are then produced on the computer screen. Ordinary least squares (ols) is used to construct the intraslice estimates of α_s and β_s. When the slider is moved, J_s is changed and the two-dimensional plot is updated. The width of J_s can be changed as necessary. The result is illustrated in Figure 3 for the ethanol data using the slice indicated in Figure 1 and without intraslice detrending; that is, with y rather than e_s. Clearly there is an intraslice trend and thus the condition d=1 is again contradicted. The button labeled "RMLTIS" (ReMove Linear Trend withIn Slice) on the plot of Figure 3 controls the option for removing intraslice trends. When this button is clicked, intraslice residuals are used, as shown in Figure 4.

The paradigm described here comes with a cost in the form of condition (2.5) which requires that e_s and $b_s^T x$ be jointly independent of $b^T x$ within each slice. We know from (2.3) that e_s is marginally independent of $b^T x$ within the slices. Condition (2.4) was imposed so that $\text{cov}(b_s^T x, b^T x \mid b^T x \in J_s)=0$. Hopefully this will be enough to insure that $b_s^T x \perp\!\!\!\perp b^T x \mid (b^T x \in J_s)$ is a good approximation. Even so, the marginal independence conditions $e_s \perp\!\!\!\perp b^T x \mid (b^T x \in J_s)$ and $b_s^T x \perp\!\!\!\perp b^T x \mid (b^T x \in J_s)$ are not sufficient to imply the joint independence as required by (2.5). We have not encountered any situations in which (2.5) clearly failed, but this is no guarantee.

Condition (2.4) requires that for each slice we first estimate $\text{cov}(x \mid b^T x \in J_s)$ and then determine b_s. These calculations may slow the procedure. As a further approximation to facilitate calculation, it might be reasonable in some cases to replace condition (2.4) with $b_s^T \text{cov}(x) b = 0$. In this way b_s is constant from slice to slice and thus needs to be determined only once at the outset. In fact, the usual moment estimate of the covariance of x is used in the calculation of b_s in Figures 3 and 4.

2.3 Overview

This section outlines the basic ideas of graphical regression when there are two predictors. The essential problem rests with determining the dimension d of the minimum dimension reduction subspace. If there is no systematic relationship evident in the rotating three-dimensional plot {y, x}, then $y \perp\!\!\!\perp x$, $S_{y|x} = S(0)$ and d=0 are indicated. If it appears that there is a sufficient two-dimensional plot $\{y, b^T x\}$, then $y \perp\!\!\!\perp x \mid b^T x$, $S_{y|x} = S(b)$ and d=1 are indicated. Otherwise, $S_{y|x}=R^2$, d=2 and the full three-dimensional plot will be needed to understand the regression structure. For the ethanol data we concluded that d=2 since we were unable to find a two-dimensional projection that seemed sufficient. After d

has been determined, we not only have the dimension of the minimum dimension reduction subspace, but we know about any curvilinear relationships and/or heteroscedasticity in the data. Various graphical methods other than those described above are available to aid in the dimension decision. Scatterplot smoothers like LOESS (See Chambers and Hastie, 1992, for example) could be used in cases where it is not immediately evident if d>0. The paradigm described here will play a central role in dealing with more than two predictors.

3. Regression with Many Predictors

The minimum dimension reduction subspace plays a central role when dealing with two predictors. The same type of construction serves to ground graphical exploration with many predictors.

3.1 Dimension Reduction Subspaces

Let η be a p×t matrix so that

$$y \perp\!\!\!\perp x \mid \eta^T x \tag{3.1}$$

Such an η always exists since (3.1) is trivially true when η=I. It is not unique, however, since if (3.1) holds then it also holds when η is replaced by any basis for $S(\eta)$. Thus (3.1) is really a statement about a subspace rather than a particular basis η. Following the rationale in Section 2, let $S_{y|x}$ denote a subspace of minimal dimension d so that (3.1) holds for any basis $\eta_{y|x}$ of $S_{y|x}$. The minimal dimension d is then the smallest number of linear combinations of x so that (3.1) holds. A *minimum dimension reduction subspace* $S_{z|w}$ for the regression of a response variable z on a vector of predictors w always exists, and we assume that all such subspaces are unique. For further discussion of uniqueness, see Cook (1994). This approach was stimulated by the work of Li (1991) who uses dimension reduction subspaces in the context of sliced inverse regression.

If $S_{y|x}$ were known then we could abandon the full (p+1)-dimensional plot {y, x} in favor of the sufficient replacement $\{y, \eta_{y|x}^T x\}$. Hopefully d will be small, say 1 or 2, since the sufficient plot could then be viewed in full. Dimensions of 3 or 4 may still represent a considerable reduction when faced with many predictors. As in the case of two predictors, the essential problem is to estimate $S_{y|x}$ graphically. Once this is done we can choose any basis $\hat{\eta}_{y|x}$ for the estimate and take the plot $\{y, \hat{\eta}_{y|x}^T x\}$ as the basic output from graphical regression, continuing the analysis as necessary to understand more of the regression structure in the estimated sufficient plot.

The graphical methods that we can employ to estimate $S_{y|x}$ must be confined to at most three-dimensional plots of the form $\{y, \delta^T x\}$, where δ is a user-selected $p \times q$ full rank matrix with $q \leq 2$, since these are the only ones that can be viewed in full. In particular, the methods of Section 2 can be used directly to estimate the minimum dimension reduction subspace $S_{y|\delta^T x}$ for $F(y|\delta^T x)$. A crucial issue is whether there are conditions under which we can choose an appropriate δ so that $S_{y|\delta^T x}$ furnishes clear information about $S_{y|x}$, at least in the population. Certain obvious possibilities can be ruled out immediately. For example, it is widely recognized that two and three-dimensional plots of y against pairs of coordinates, $\{y, (x_j, x_k)\}$, do not necessarily give reliable information about the form of the full regression (See, for example, Chambers, Cleveland, Kleiner, and Tukey 1983, p. 268). It is not difficult to see as well that $\{y, (x_j, x_k)\}$ will not necessarily give clear information on $S_{y|x}$.

It seems that strong nonlinear relationships among the predictors is what keeps us from constructing low dimensional plots $\{y, \delta^T x\}$ in which $S_{y|\delta^T x}$ has a clean relationship with $S_{y|x}$. In the next section we describe how some progress can be made when the predictors follow a multivariate normal distribution. This assumption rules out the possibility of nonlinear relationships among the predictors but allows for strong linear relationships. We indicate how the results can be extended beyond the normal case in Section 3.3.

3.2 Normal Predictors

In this section we assume that x follows a nonsingular multivariate normal distribution. Partition $x^T=(x_1, x_2)$ where x_2 is $q \times 1$, $q \leq 2$, conformably partition a $p \times d$ basis for $S_{y|x}$, $\eta_{y|x}^T = (\eta_1^T, \eta_2^T)$ and rewrite (3.1) as

$$y \perp\!\!\!\perp x \mid (\eta_1^T x_1, \eta_2^T x_2) \tag{3.2}$$

The general approach of this section is to first estimate various *component subspaces* $S(\eta_2)$ which are at most two-dimensional and then construct an estimate of $S_{y|x}$ by using direct sums of the estimated component subspaces.

Let $e_{2|1} = x_2 - E(x_2|x_1)$ denote the $q \times 1$ vector of residuals from the indicated population regression of x_2 on x_1. These residuals correspond to the linear combination $\delta^T x$ mentioned above, the matrix δ being selected so that $e_{2|1} = \delta^T x$. When considering sample versions a bit later, ols can be used to estimate the regression function $E(x_2|x_1)$ since x is normal. The minimum dimension reduction subspace for $F(y|e_{2|1})$ will be denoted by $S_{y|e_{2|1}}$.

Lemma 3.1. If x is normally distributed and $e_{2|1}$ and η_2 are as defined above then

$$\text{(a) } y \perp\!\!\!\perp e_{2|1} \mid \eta_2^T e_{2|1} \text{ and (b) } S_{y|e_{2|1}} \subset S(\eta_2) \tag{3.3}$$

Justification: Part (a) follows from Cook (1994, Lemma 4.1); part (b) follows immediately from part (a).

The results in Lemma 3.1 are useful because they say that the minimum dimension reduction subspace for the regression of y on $e_{2|1}$ is contained within $S(\eta_2)$ which is a component of $S_{y|x}$. However, because we are not guaranteed that $S_{y|e_{2|1}} = S(\eta_2)$, the two or three-dimensional plot $\{y, e_{2|1}\}$ may miss relevant information about $S(\eta_2)$. This possibility can be ruled out with an assumption:

Lemma 3.2 Let $\gamma_{2|1}$ denote a basis for $S_{y|e_{2|1}}$ and assume that x is normally distributed. Then $S_{y|e_{2|1}} = S(\eta_2)$ if and only if

$$(y, x_1) \perp\!\!\!\perp e_{2|1} \mid \gamma_{2|1}^T e_{2|1} \tag{3.4}$$

Justification: The result follows from Lemma 4.2 of Cook (1994) or, starting somewhat farther back, from Lemmas 4.1-4.3 of Dawid (1979).

Since x is normally distributed, $x_1 \perp\!\!\!\perp e_{2|1}$ and it follows from the construction of $\gamma_{2|1}$ that $y \perp\!\!\!\perp e_{2|1} \mid \gamma_{2|1}^T e_{2|1}$. Thus, the marginal independence conditions in (3.4) hold.

However, as in the discussion of condition (2.5), marginal independence does not necessarily imply joint independence. Nevertheless, we feel that (3.4) will be a reasonable assumption in many applications since some rather extreme interactions must be present otherwise.

Lemma 3.2 is important because it gives conditions for low dimensional plots of the form $\{y, e_{2|1}\}$ to provide clean information about $S_{y|x}$ via the component subspace $S(\eta_2)$. To explore the implications of this, we consider the two and three-dimensional plots separately, in each case assuming that (3.4) holds.

3.2.1 Two-dimensional Plots, q=1. In this case x_2 is a single predictor, η_2 is a 1×d vector and thus $S(\eta_2)$ equals either S(0) or R^1. Since $S_{y|e_{2|1}} = S(\eta_2)$ this means that $S_{y|e_{2|1}}$ is either S(0) or R^1. If there is no dependence evident in a sample version of the two-dimensional plot $\{y, e_{2|1}\}$ then the conclusion $S(\eta_2)$=S(0) is indicated. This in turn

implies that $y \perp\!\!\!\perp x \mid \eta_1^T x_1$ and thus that x_2 is not needed in the full regression. On the other hand, if a sample version of $\{y, e_{2|1}\}$ does exhibit some dependence then $S(\eta_2)=R^1$ is indicated and x_2 is needed in the full regression.

By considering all p version of the two-dimensional plot $\{y, e_{2|1}\}$ as x_2 is set to each predictor in turn, we are thus able to determine if each predictor is needed in the full population regression. As a practical matter, our ability to detect dependence visually will be limited by sample size and the strength of any dependence. Scatterplot smoothers for location and spread may be useful here.

We usually begin in practice by viewing sample versions of all p plots $\{y, e_{2|1}\}$. If it is clear that there is a plot which exhibits no dependence, then we may delete the corresponding variable and recompute the remaining plots as a way of reducing variation.

3.2.2 Three-dimensional Plots. The paradigm described in Section 2 can be used to analyze three-dimensional plots $\{y, e_{2|1}\}$ arising when selecting a subset x_2 of q=2 predictors. We again rely on the identity $S_{y|e_{2|1}} = S(\eta_2)$ to reach one of 3 possible decisions: Either $S(\eta_2) = S(0)$, $S(\eta_2) = S(b)$ for some 2×1 vector b, or $S(\eta_2) = R^2$.

The conclusion $S(\eta_2) = S(0)$ is indicated when there is no dependence evident in a sample version of $\{y, e_{2|1}\}$. In this case the predictors x_2 might be deleted from the analysis, as in the previous section. Three-dimensional plots are generally superior to two-dimensional plot when deciding if $\dim(S(\eta_2))>0$ since the variation around any systematic trends will be less in a three-dimensional plot: $\text{var}[y - E(y|e_{2|1})] \le \text{var}[y - E(y|a^T e_{2|1})]$ for any a. In practice, if the two-dimensional plots of the previous section lead to the tentative conclusion that two of the predictors are not needed in the full regression, then we usually view the corresponding three-dimensional plot before making the final determination.

The conclusion $S(\eta_2) = S(b)$ is indicated when a sample version of $\{y, e_{2|1}\}$ exhibits a 1-dimensional structure, as described in Section 2. The vector b gives the corresponding linear combination $b^T x_2$ from the visual fit. The implication is that the two predictors in x_2 can be replaced by the single predictor $b^T x_2$ and that $\{y, (x_1, b^T x_2)\}$ is a sufficient replacement for $\{y, x\}$.

The conclusion $S(\eta_2) = R^2$ is indicated when a sample version of $\{y, e_{2|1}\}$ exhibits 2-dimensional structure, again as described in Section 2. The implication is that both predictors in x_2 are needed in the full regression and that the situation is relatively complicated so they cannot be replaced by a single linear combination. The only recourse

in such cases is to select a different pair of predictors in the hope of achieving further dimension reduction.

Once $S(\eta_2)$ has been characterized in terms of one of these three possible decisions and the appropriate action has been taken, we can choose a new pair of variables and begin again. This leads to a sequential procedure which may eventually produce a useful characterization of $S_{y|x}$.

3.2.3 Example. To illustrate some of the characteristics of a sequential procedure for graphically estimating $S_{y|x}$, consider a regression problem with p = 4 predictors $w^T=(w_1,...,w_4)$, d = 2 and $S_{y|x}$ spanned by the columns of

$$\eta_{y|x} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{3.5}$$

We continue to use x_1 and x_2 to describe a partition of w as in Sections 3.2.1 and 3.2.2.

Consider first inspecting all possible two-dimensional plots $\{y, e_{2|1}\}$ as described in Section 3.2.1. From (3.5) we see that $S(\eta_2) = R^1$ for the first 3 predictors and $S(\eta_2) = S(0)$ for w_4. Thus the plots $\{y, e_{2|1}\}$ for the first 3 predictors should give information to contradict the condition $y \perp\!\!\!\perp e_{2|1}$, but for w_4 there should be no such information which implies that $y \perp\!\!\!\perp w_4 \mid (w_1, w_2, w_3)$ and thus that w_4 can be deleted from the regression without loss.

With w_4 deleted from the analysis, we now turn to three-dimensional plots to see if further reduction is possible. If $\eta_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, corresponding to $x_1=w_1$ and $x_2^T=(w_2, w_3)$, then no reduction is possible since $S(\eta_2)=R^2$ and thus the three-dimensional plot $\{y, e_{2|1}\}$ should exhibit 2-dimensional structure. The same conclusion applies if we set $x_1=w_2$ and $x_2^T=(w_1, w_3)$. However, dimension reduction is possible when $x_1=w_3$ and $x_2^T=(w_1, w_2)$ since in this case $\eta_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $S(\eta_2) = S((1, 1)^T)$. The method of analysis described in Section 2 will lead to the conclusion that w_1 and w_2 can be replaced by their sum, at least theoretically.

With w_4 deleted from the analysis and with w_1 and w_2 replaced by their sum, we have a new regression problem with 2 predictors, w_1+w_2 and w_3. From (3.5) it is easily seen that the minimum dimension reduction subspace for this new regression is R^2 and thus no further dimension reduction is possible. The final graphical regression plot is then $\{y, (w_1+w_2, w_3)\}$.

Consider next a second version of this example, this time with

$$\eta_{y|x} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.6}$$

No reduction is possible with two-dimensional plots since all variables are needed in the regression. Similarly no reduction is possible with three-dimensional plots $\{y, e_{2|1}\}$ since $S(\eta_2)=R^2$ for all possible choices of η_2. We were able to achieve dimension reduction under (3.5) because there are choices for η_2 with $\dim(S(\eta_2))<2$, but this is not possible under (3.6). The only way to achieve dimension reduction under (3.6) with the present paradigm is to use four-dimensional plots $\{y, e_{2|1}\}$ in which x_2 consists of three predictors. Although we can imagine how it should be done, we have yet to construct a successful interface for the analysis of four-dimensional plots.

To guarantee the ability to reduce x to $d=\dim(S_{y|x})$ linear combinations, x_2 must consist of d+1 predictors which results in a (d+2)-dimensional plot $\{y, e_{2|1}\}$. The paradigm described here always allows determination of $S_{y|x}$ when d=1. Depending on $S_{y|x}$ it may also allow dimension reduction when d>1, as illustrated with (3.5), but there are no longer any guarantees, as illustrated with (3.6).

3.2.4 Adding an Objective Function. Situations in which d=2 and $\dim(S(\eta_2))=2$ for all possible choices of the 2×1 vector x_2 really push the limits of what is possible with the graphical regression paradigm presented so far. This is an annoying limitation because, as long as we are restricted to at most three-dimensional plots, dimension reduction cannot be guaranteed theoretically when d=2. The possibility of using four-dimensional plots seems remote. However, some progress may be possible if we allow strategic use of an objective function.

Consider summarizing the data through the fit of a linear predictor $a + b^Tx$ via an objective function L(g, y) which is convex in its first argument g,

$$(\hat{a}, \hat{b}) = \arg\min \tfrac{1}{n} \Sigma_i L(a + b^Tx_i, y_i) \tag{3.7}$$

The function L might be chosen to yield ols estimates or Huber's M-estimates, for example. There is no assumption here that the linear predictor is suitable or that it even yields a sensible fit to the data. Let

$$(\alpha, \beta) = \arg\min E[L(a + b^Tx, y)] \tag{3.8}$$

denote the population version of (3.7), and assume that β is unique. The expectation in (3.8) is computed with respect the joint distribution of y and x, and is assumed to be finite. It follows as a slight extension of Li and Duan (1989, Theorems 2.1 and 5.1) that $\hat{b}$

converges almost surely to β and that $\beta \in S_{y|x}$. Assuming that $\beta \neq 0$, this result may help sort out graphical regression problems in which d=2 and yet all three-dimensional plots $\{y, e_{2|1}\}$ exhibit 2-dimensional structure.

We now use the vector $\beta \in S_{y|x}$ to construct a set of linearly transformed predictors $w = B^T x$ where $B=(\beta, \beta^*)$ and β^* is any $p \times (p-1)$ matrix that extends β to a basis for R^p. A basis for $S_{y|w}$, the minimum dimension reduction subspace for the regression of y on w, is then

$$\eta_{y|w} = \left(\binom{1}{0} | \beta_1\right)$$

for some $p \times 1$ vector β_1. Subspaces of this form always allow reduction to two linear combinations of the predictors w with three-dimensional plots. Once found, such subspaces can be backtransformed easily to yield the corresponding linear combinations of x. We will need to replace β by its estimate $\hat{b}$ from (3.7) in practice.

Allowing the use of an objective function as in (3.8) theoretically guarantees reduction to the appropriate linear combinations of x when d=2 and the conditions given above are met. The procedure may often help when d>2, but again there are no guarantees.

3.2.5 Reducing Variation. All of the variation present in y enters into the two and three-dimensional plots $\{y, e_{2|1}\}$ used to determine $S(\eta_2)$. It turns out that it is possible to reduce the variation in these plots without loss of information by replacing y with a set of residuals from a regression of y on x_1. Reducing the variation is potentially useful because it may make $S(\eta_2)$ easier to identify visually.

Let $r_{y|1}$ denote a typical population residual from some regression of y on x_1. There is no requirement that the associated model be an adequate description of $F(y|x_1)$, although there are obvious advantages if this is the case. In practice, the residuals could be from an ols fit of y on x_1, from a fit of a full second-order quadratic model, or from the fit of a generalized additive model, for example.

Lemma 3.3. Assuming that x is normally distributed and with the notation established above,

$$\text{(a) } r_{y|1} \perp\!\!\!\perp e_{2|1} \mid \eta_2^T e_{2|1} \text{ and (b) } S_{r_{y|1}|e_{2|1}} \subset S(\eta_2)$$

where $S_{r_{y|1}|e_{2|1}}$ is the minimum dimension reduction subspace for the regression of $r_{y|1}$ on $e_{2|1}$.

Justification: It is not difficult to see that $y \perp\!\!\!\perp e_{2|1} \mid (\eta_2^T e_{2|1}, x_1)$. This plus the condition $x_1 \perp\!\!\!\perp e_{2|1}$ implies that $(y, x_1) \perp\!\!\!\perp e_{2|1} \mid \eta_2^T e_{2|1}$ and part (a) follows. Part (b) follows immediately from (a).

Lemma 3.3 is the counterpart of Lemma 3.1 with y replaced by $r_{y|1}$. It shows that the minimum dimension reduction subspace for the regression of $r_{y|1}$ on $e_{2|1}$ is always contained in the desired component subspace $S(\eta_2)$. Further, by adopting a condition for the regression of $r_{y|1}$ on $e_{2|1}$ similar to (3.4), we can again arrive at the desired conclusion $S_{r_{y|1}|e_{2|1}} = S(\eta_2)$.

For use in practice, consider taking $r_{y|1}$ to be the residuals from the ols regression of y on x_1. Further, construct a sample version of $e_{2|1}$ by using ols to estimate the components of $E(x_2|x_1)$. Then the sample version of the plot $\{r_{y|1}, e_{2|1}\}$ is just a two or three-dimensional added variable plot as discussed in Cook and Weisberg (1989). We use added variable plots for determining the dimension of component subspaces in the examples of Section 4.

3.3 A Little Beyond Normality

The essential results given above extend fairly easily to other distributions that force linear relationships among the predictors. In particular, if we assume that $E(x|Ax)$ is finite and linear in Ax for all $t \times p$ matrices A, then it follows from Eaton (1986) that x must have an elliptically contoured distribution.

Assume then that x has an elliptically contoured distribution with mean μ and covariance matrix Σ, and partition $\Sigma = (\Sigma_{jk})$ according to the partitioning of x, j,k=1,2. Next, define $\Sigma_{2|1} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$ and $\omega(e_{2|1}) = e_{2|1}^T \Sigma_{2|1}^{-1} e_{2|1}$, where $e_{2|1} = x_2 - E(x_2|x_1)$ as in the normal case. Then it is not difficult to verify that

$$y \perp\!\!\!\perp e_{2|1} \mid (\eta_2^T e_{2|1}, \omega(e_{2|1})) \tag{3.9}$$

The difference between this and the normal case described in Lemma 3.1 is the presence of the term $\omega(e_{2|1})$ which reflects the length of $e_{2|1}$ relative to the contours of $\Sigma_{2|1}^{-1}$. Because of this it is at least theoretically possible to see systematic patterns in the plot $\{y, e_{2|1}\}$ that vary across the contours of $\omega(e_{2|1})$ in addition to any effects arising from F(y|x). We have not found this possibility worrisome in practice since the systematic trends due to $\omega(e_{2|1})$ are usually small relative to background variation.

Nevertheless, following Cook (1994) the dependence of (3.9) on $\omega(e_{2|1})$ can be removed by standardizing $e_{2|1}$: Let $e^*_{2|1} = e_{2|1}/(\omega(e_{2|1}))^{1/2}$. Then

$$y \perp\!\!\!\perp e^*_{2|1} \mid \eta_2^T e_{2|1} \tag{3.10}$$

and $\{y, e^*_{2|1}\}$ is the corresponding normalized plot for determining $S(\eta_2)$ visually.

Regardless of the specific distributional assumptions for x, we expect that the paradigm outlined in Section 3.2 will give reasonable results in practice as long as there is no strong nonlinear dependence among the predictors. In particular, the results of Section 2.3.4 hold for elliptically contoured predictors and (3.10) can be modified for variance reduction as in Section 2.3.5.

4. Examples

4.1 Constructed Data

We first use a constructed data set to illustrate the application of graphical regression methods when p>2. The data consists of n=150 observations generated according to the model $y = w_1(w_1 + w_2 + 1) + \varepsilon$, where the p = 3 covariates $w^T = (w_1, w_2, w_3)$ and the error ε are independent standard normal random variables. The third covariate w_3 was not used in the generation of y. The minimum dimension reduction subspace for the regression of y on w is spanned by the pair of vectors $(1,0,0)^T$ and $(1,1,0)^T$.

According to the ideas in Section 3.2.5, an appropriate tool for graphical regression analysis is the added variable plot. All estimates necessary for the construction of the plots in this example were obtained via ols. We begin with the added variable plot $\{e_{y|3}, (e_{1|3}, e_{2|3})\}$ for $x_2 = (w_1, w_2)$ after $x_1 = w_3$. The population version of this plot will show a 2-dimension structure. This 2-dimensional structure is somewhat clear in our sample as we rotate the three-dimensional added variable plot, but the slicing idea developed in Section 2.2 makes seeing the second dimension easier. Figure 5 presents the projection $\{e_{y|3}, 0.95e_{1|3} + 0.30e_{2|3}\}$ that we consider to be the best. If a 1-dimensional structure were sufficient to describe $\{e_{y|3}, (e_{1|3}, e_{2|3})\}$, then the points contained in any slice parallel to the y-axis would be independent of w. Figure 6 contains a plot of selected points of Figure 5, with the same vertical axis, but with the horizontal axis in the direction orthogonal to (0.95, 0.30) relative to the inner product determined by the sample covariance matrix for $(e_{1|3}, e_{2|3})$. This figure shows a clear linear tendency that persists across several slices. After trying a few other linear combinations and consistently finding dependence in the

orthogonal direction, we correctly conclude that this plot exhibits a 2-dimensional structure and we make no dimension reduction.

We continue the analysis by considering next the added variable plot $\{e_{y|1}, (e_{2|1}, e_{3|1})\}$ for $x_2 = (w_2, w_3)$ after $x_1 = w_1$. The population version of this plot would exhibit a 1-dimensional structure with $\eta_2 = (1, 0)^T$ spanning the subspace. Figure 7 shows the sample added variable plot in the $(1, 0)^T$ direction. It may be hard to see any interesting pattern. However, we know from the analysis of Figure 5 that the dependence of y on w_1 is nonlinear, so that the ordinary residuals $e_{y|1}$ may not remove sufficient variation along the vertical axis to see the trend. The method of Section 3.2.5 lets us use any residuals from y on w_1 as the vertical axis in this plot. Figure 8 shows a plot of y versus w_1 smoothed with a triangular kernel. We will use the residuals from this fit as the vertical axis. (In our implementation, the GREG Methods button allows us to change the vertical axis to residuals from a smoother, or the residuals from a full second-order quadratic model for y on x_1.) The resulting plot is shown here in Figure 9. The double fan-shaped pattern is a result of the w_1w_2 interaction. We would conclude the investigation of the plot $\{e_{y|1}, (e_{2|1}, e_{3|1})\}$ by stating that it exhibits a 1-dimensional structure in the $(1, 0)^T$ direction.

We have now reduced the number of predictors in this problem from three to two. The graphical analysis would be concluded by an investigation of the three-dimensional plot $\{y, (w_1, w_2)\}$. This will be essentially equivalent to the investigation of the added variable plot for (w_1, w_2) after w_3, which was discussed in the first paragraph of this section. Thus, our final conclusion is that we have $d = 2$ and the minimum dimension reduction subspace is spanned by the vectors $(1,0,0)^T$ and $(0,1,0)^T$.

This example has shown another application of the slicing technique, and the variance reduction ideas of Section 3.2.5, but has also illustrated the idea of dimension reduction and how that may take place in practice. In the next example we illustrate that dimension reduction is possible with a large data set and many predictors.

4.2 Environmental Contamination

A large simulation model was developed to aid in a study of an environmental contaminant introduced into an aquatic ecosystem. A good appreciation of the ecological risk associated with contamination requires an understanding of the long-term fate of a contaminant and how it filters through the ecosystem. The environmental model is based on the assumption that the contaminant first becomes available in the ecosystem by dissolution to water and then moves through the food web via one organism consuming another. It represents the ecosystem as 11 compartments of the food web -- water,

sedimented dead organic matter, phytoplankton and carnivorous fish, for example -- that can receive and pass along the contaminant. Four other compartments serve as sinks that only receive the contaminant.

The response y is the steady state concentration of the contaminant in sedimented dead organic matter. There are 9 predictors, w_1, ..., w_9; w_1 is the initial concentration of the contaminant in water and w_2-w_9 represent various transfer rates between compartments. We are interested in seeing what graphical regression has to offer for the development of a relatively simple model based on about 1000 runs of the simulation model. As a result of ecological considerations and initial investigation of bivariate relationships in a scatterplot matrix, we begin by taking the logarithms of the response y and the predictors w_1, w_3 and w_4.

The sequential graphical regression process requires analyzing a three-dimensional added variable plot at each step, making a dimension decision and reducing the number of predictors if possible. We have found it useful to initially order the variables roughly on the strength of relationship indicated in a series of two-dimensional added variable plots. The construction of three-dimensional added variable plots can then begin with the variables displaying the strongest or weakest relationships in the two-dimensional plots.

A quick look at the two-dimensional added variable plots suggests that the predictors with the most predictive power are $v^T = (\log(w_1), w_2, \log(w_3), \log(w_4))$ and that the remaining predictors are relatively unimportant. We begin the construction of three-dimensional added variable plots with w_5 - w_9, the predictors displaying no clear relationship in the two-dimensional plots. As discussed in Section 3.2.2, the impression that these predictors are not needed for the regression should be corroborated by viewing the corresponding three-dimensional plots. We first viewed the added variable plot for x_2 = (w_5, w_6) after $x_1 = (\log(w_1), w_2, \log(w_3), \log(w_4), w_7, w_8, w_9)$. Rotation of this plot showed no systematic patterns and we concluded that the plot exhibits a 0-dimensional structure. This eliminates w_5 and w_6 from the model. Proceeding sequentially, we also eliminated w_7-w_9 since the added variable plots for w_7 and w_8 after (v, w_9) and for w_9 after v seem to be 0-dimensional as well. We now have only the 4 predictors in v remaining. We reached the same conclusion when considering the predictors w_5 - w_9 in various orders.

Figure 10 gives the added variable plot for $(\log(w_3), \log(w_4))$ after $(w_2, \log(w_1))$. Rotation of this plot shows a dominant linear trend. There is curvature in a second direction, however, since close visual inspection will reveal that the 3-dimensional structure resembles a tilted trough. In order to take a closer look at the 2-dimensional structure, we

again use the method discussed in Section 2.2. Figure 11 shows quadratic trends in the four separate slices of the added variable plot in Figure 10. The symbol used in each of these four plots is consistent with the symbols in Figure 10. This is clear evidence of a 2-dimensional structure and thus we can not use this plot to reduce the dimension of the problem further. We now know that $\dim(S_{\log(y)|w}) \geq 2$.

Investigation of the added variable plots for $(\log(w_1), \log(w_3))$ after $(w_2, \log(w_4))$ and $(\log(w_1), \log(w_4))$ after $(w_2, \log(w_3))$ also exhibit 2-dimensional structure, but not quite as clearly as Figure 11. The added variable plot for $(\log(w_1), w_2)$ after $(\log(w_3), \log(w_4))$ exhibits a 1-dimensional structure, however. Figure 12 shows our sufficient reduction of this plot with the horizontal axis rotated to $z = 0.01\log(w_1) - 0.99\, w_2$. Thus, we have now reduced the problem to three predictors, z, $\log(w_3)$ and $\log(w_4)$. Each of the three possible three-dimensional added variable plots with these predictors shows a 2-dimensional structure.

Our conclusion is that the dependence of log(y) on the original 10 predictors can be reduced to a dependence on three predictors $\log(w_3)$, $\log(w_4)$ and z, and that $\dim(S_{\log(y)|w}) = 2$ or 3. We may have a situation similar to the example corresponding to equation (3.6), however. We continued the analysis by applying the ideas of Section 3.2.4 to the final three predictors, and concluded that the dimension is 2. Our final estimate of a basis for $S_{\log(y)|w}$ is

$$\hat{\eta}_{\log(y)|w} = (\eta_1, \eta_2) = \begin{pmatrix} 0.98 & 0 \\ -89.52 & 0 \\ 0.39 & -0.73 \\ 0.43 & 0.68 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{4.1}$$

where the predictors are taken in the order of their subscripts, $w^T = (\log(w_1), w_2, \log(w_3), \ldots, w_9)$.

The plot $\{\log(y), (\eta_1^T w, \eta_2^T w)\}$ shows the same tilted trough structure as described in connection with Figure 10, with $\eta_1^T w$ being the linear direction. Fitting the quadratic linear model $\log(y) = b_0 + b_1(\eta_1^T w) + b_2(\eta_2^T w) + b_{22}(\eta_2^T w)^2$ by ols yields $R^2 = 0.90$, so graphical regression has been able to explain much of the variation in y.

Recently developed methods like SIR and pHd (Li 1991, 1992) can be used to estimate $S_{\log(y)|w}$ without the need for intermediate graphics. Application of SIR to the data of this example yields only a single significant direction and thus an estimate of $S_{\log(y)|w}$ with dimension 1. The significant direction found by SIR is close to the subspace spanned by (4.1): The angle between it and it's projection onto this subspace is 6.48 degrees. The

failure of SIR to find a two-dimensional subspace is not really surprising since SIR can not find quadratics very well; pHd would surely do much better in this regard. In any event, we find that graphical regression can often complement other methods of analysis.

5. Rounding Out the Ideas

In this section we consider a number of issues to round out the ideas and to provide some connection with fairly standard methods used in data analysis. These issues were neglected in the previous discussion to avoid interrupting the development of the central ideas.

5.1 Discrete Responses

Essentially no assumptions have been imposed on the response variable other than the univariate requirement. In particular, all of the theoretical results hold for both continuous and discrete responses. The convex loss function L used in Section 3.2.4, for example, can be derived from the log likelihood for a generalized linear model. Nevertheless, for some discrete responses the theory can be quite difficult if not impossible to apply without additional aids.

For instance, assume that y|x is a Bernoulli random variable taking the values 0 or 1 with probability depending on x. Application of the methods in Section 3.2 requires visually extracting an estimate of the minimum dimension reduction subspace in two or three-dimensional plots of the form $\{y, e_{2|1}\}$. A two-dimensional plot will consist of two horizontal lines of points, one at y=0 and one at y=1. While it may be possible to see dependence in such plots, the required visual analysis will usually be much more difficult than that for continuous responses. The introduction of scatterplot smoothers seems essential if there is to be a hope of success in practice.

The situation seems even more problematic when $\{y, e_{2|1}\}$ is three dimensional. In that case the plot will consists of two parallel planes of points at y=0 and y=1. Deciding on visual inspection if the plot has 0-dimensional structure seems very difficult, to say nothing about the elusive task of deciding between 1 and 2-dimensional structure. Again the introduction of smoothers seems essential. Some ideas for smoothing plots of binary responses is available in Fowlkes (1987). More investigation is clearly needed to see if there is any practical potential to the idea of graphical binary regression.

5.2 Transformations

Monotonic transformations of the response can be used without disturbing any of the structure underlying a graphical regression problem since the minimum dimension reduction subspace is invariant under such transformations: If t(y) is a monotonic transformation of y then clearly $S_{t(y)|x} = S_{y|x}$. Response transformations in graphical regression can be used to simplify the nature of the dependence of y on x, just as they are used in parametric modeling. Suppose for example that d=1 and that $S_{y|x}=S(b)$. Then it may happen that $E(y|b^Tx)$ is a nonlinear function of b^Tx and that $var(y|b^Tx)$ depends on x as well. In such cases an appropriate transformation t(y) will often have the desirable properties that $E(t(y)|b^Tx)$ is essentially linear in b^Tx and $var(t(y)|b^Tx)$ is essentially constant. This is the rationale for taking the log of the response in the example of Section 4.2.

Transformations of the predictors is another story. Let $t(x) = (t_k(x_k))$, k=1,..,p, denote a vector-valued transformation of x. One goal might be to choose t so that

$$\dim(S_{y|t(x)}) < \dim(S_{y|x}) \qquad (5.1)$$

and thus possibly reduce the complexity of the regression problem by reducing dimension. Suppose, for example, that $y \perp\!\!\!\perp x \mid \|x\|$ so that $\dim(S_{y|x})=p$. As long as we stay with linear combinations of the original predictors x, no dimension reduction is possible in this case. However, if we set $t(x)=(x_k^2)$ then (5.1) becomes $1=\dim(S_{y|t(x)}) < \dim(S_{y|x})=p$, resulting in considerable dimension reduction. A conflict may occur between (5.1) and the requirement that there be no strong nonlinear relationships among the predictors. Fortunately, we may be able to achieve both goals simultaneously in practice, as seems to be the case in the second example of Section 4. Turning to nonparametric regression, there is a notable connection between (5.1) and generalized additive modeling which may proceed based on the assumption that there exists a predictor transformation t such that $\dim(S_{y|t(x)})=1$. Methods for achieving (5.1) in the context of graphical regression are under study.

For nonsingular linear transformations $t(x) = A^Tx$, the minimum dimension reduction subspaces are related by $\eta_{y|t(x)} = A^{-1}\eta_{y|x}$ and consequently $\dim(S_{y|t(x)}) = \dim(S_{y|x})$. Further, it is not difficult to verify that the graphical regression procedure outlined here is invariant under nonsingular linear transformations, although that may not be the case in practice depending on the consistency of our visual interpretation of various plots. These facts are helpful when the two horizontal variables in a three-dimensional plot for determining a component subspace are highly correlated. In such cases we simply orthogonalize the variables on the horizontal axes, estimate the corresponding component subspace in the transformed coordinates and then backtransform to the original coordinates.

See Cook and Weisberg (1990) for additional comments on orthogonalization in three-dimensional plots.

5.3 Model Diagnostics

One important practical benefit that may come of this development is the ability to construct a comprehensive series of diagnostic plots for an empirical regression model. Suppose that we have conducted a careful regression analysis arriving at an estimated model and a set of residuals $\hat{r}$. The modeling process could involve transformations of the response or predictors, fitting a generalized additive model, the addition of quadratic or cross product terms in the original predictors x, or any of the available regression methods. If the model is a good representation of the data then the residuals $\hat{r}$ should *appear* as an independent and identically distributed sample. Strictly speaking, the residuals can not be independent and identically distributed because of the second-order effects introduced by substituting estimates for unknown parameters. Nevertheless, such effects are usually small and well-understood, particularly in the context of linear regression (Cook and Weisberg 1982, Cox and Snell 1968). We ignore such effects in this section.

Let $S_{\hat{r}|x}$ denote the minimum dimension reduction subspace for the regression of $\hat{r}$ on x. If the developed model is a good reflection of the data, then we should have $S_{\hat{r}|x}=(0)$, implying that $\hat{r} \perp\!\!\!\perp x$. Otherwise the model is deficient and remedial action may be necessary. When there are no strong nonlinear trends among the predictors, the graphical regression procedure described in Section 3 essentially guarantees the detection of any model deficiencies up to our ability to see systematic patterns. Failing to find any deficiencies we may have good reassurance that the model is reasonable. This seems to be a significant advance beyond the standard diagnostic residual plots.

The procedure may be carried a bit further to indicate something about the necessary remedial actions when deficiencies are found. Suppose we conclude that a particular three-dimensional plot $\{\hat{r}, e_{2|1}\}$ exhibits a 1-dimensional structure with visually fitted direction $b^T e_{2|1}$. This will be a good indication that the model is deficient in the form of the linear combination $b^T x_2$ and thus that we should reconsider the way in which x_2 enters the model.

5.4 Another Objective Function

In Section 3.2.4 we introduced an objective function to help in situations where d=2 and yet $\dim(S(\eta_2) = 2$ for all choices of the 2×2 matrix η_2. An objective function may also help in the intermediate analysis of three-dimensional plots $\{y, e_{2|1}\}$ or $\{r_{y|1}, e_{2|1}\}$. As in (3.8), let

$$(\alpha_2, \beta_2) = \arg\min E[L(a + b^T e_{2|1}, y)]$$

where L is as generally described in Section 3.2.4. If x is an elliptically contoured random variable then $e_{2|1}$ is elliptically contoured. Using Lemma 3.2 and the rationale of Section 3.2.4 it then follows that $\beta_2 \in S_{y|e_{2|1}} = S(\eta_2)$. This result can be used to provide a baseline for visually fitting in the plot $\{y, e_{2|1}\}$.

Let $\hat{b}_2$ denote an estimate of β_2 and consider a plot where it is clear that $\dim(S(\eta_2))>0$. To help decide if $\dim(S(\eta_2)) = 1$ or 2, rotate to the two-dimensional projection $\{y, \hat{b}_2^T e_{2|1}\}$. If the data do not contradict this plot as a sufficient replacement for $\{y, e_{2|1}\}$ and it seems consistent with our visual fit, then it would seem reasonable to take $\hat{b}_2$ as an estimated basis for $S(\eta_2)$.

To see how this might be used in application, let L corresponds to the ols objective function and let $\hat{b}$ denote the ols estimate from the regression of y on x using (3.7). Use ols estimates in the construction of the sample version of $e_{2|1}$ as well. Suppose that at every sample version of the three-dimensional plot $\{y, e_{2|1}\}$ encountered in a full sequential application of graphical regression we conclude that $\dim(S(\eta_2)) = 1$ and then we use the ols estimate $\hat{b}_2$ as the estimated basis. The end result of this procedure is that $S(\hat{b})$ is the estimate of $S_{y|x}$, with the implication that the standard two-dimensional plot of responses versus the ols fitted values $\{y, \hat{b}^T x\}$ is a sufficient replacement for the full (p+1)-dimensional plot $\{y, x\}$. This conclusion follows because the ols coefficients $\hat{b}_2$ from the regression of y on the ols sample version of $e_{2|1}$ are the same as the ols estimates of x_2 in the full regression.

5.5 Outliers and Influence

Outlying observations in the response may be relatively easy to deal with in this setting since they often stand apart in multiple three-dimensional plots $\{y, e_{2|1}\}$. The *XLISP-STAT* code that we use allows for points selected in any plot to be easily deleted from all calculations. In addition it is fairly easy to mentally neglect outlying responses during visual fitting so their influence can be minimized.

Cook (1986) demonstrates that added variable plots are good graphical tools for assessing the local influence of cases in regression problems. As indicated in Section 3.2.5, the plot $\{r_{y|1}, e_{y|1}\}$ is a three-dimensional added variable plot for adding x_2 after x_1 (Cook 1987, Cook and Weisberg 1989) when $r_{y|1}$ is constructed as the ordinary residual from the ols regression of y on x_1. Thus we expect that these plots will be useful for

visually assessing influence, in addition to the variance reduction properties described in Section 3.2.5.

5.6 Assumptions on the Predictors

The main technical constraint on graphical regression is the condition that there are no strong nonlinear relationships among the predictors. This condition is similar to those required for SIR and pHd, innovative inverse regression methods recently developed by Li (1991, 1992). While the condition is a notable limitation, there are ways to mitigate its effects and thus extend applicability a bit.

One way is to transform the predictors in the hope of getting closer to an elliptically contoured distribution, at least in marginal plots $\{x_j, x_k\}$. This is part of the rationale for transforming the predictors in the example of Section 4.2. Transforming the predictors may add dimensions, but fortunately just the opposite seems to happen often in practice.

Another possibility is to remove a small fraction of the data so that the empirical cdf constructed from the remaining observed values of the predictors closely matches the cdf of a selected elliptically contoured distribution. One specific method for doing this is studied by Cook and Nachtsheim (1992). The observations removed are not intended to remain so permanently, but can be reinstated for application of any method that does not rely on elliptically contoured predictors. In addition, the response plays no role in determining which observations to remove so that the distribution of y|x is not disturbed at the values of x corresponding to the remaining data.

6. Final Remarks

It has been generally recognized for some time that characteristics of a full regression of y on x are not necessarily preserved when considering subset regressions. This investigation indicates that, while the form of the functional dependence may not be preserved, it is possible to preserve components of minimum dimension reduction subspaces. In this way certain classes of low dimensional plots have specific characteristics that relate directly to component subspaces of the full regression. The role that added variable plots play in this investigation appears to be novel.

In more traditional regression modeling, the conditional distribution of y given x is the primary focus of attention. Most of the common regression assumptions -- linearity, homoscedasticity, and normality, for example -- are imposed on y|x and generally facilitate the analysis. There is a large body of diagnostic and remedial methodology for detecting when such assumptions fail and for molding problems to fit them. Except for concerns like

influence, leverage and collinearity, the distribution of the predictors generally plays a minor role in traditional regression modeling.

Nothing in statistics comes free. In the present investigation, the marginal distribution of the predictors is the recipient of the assumptions while y|x is essentially unconstrained. We have rarely found the necessary assumptions on x to seriously fail. Nevertheless, as in traditional regression modeling, it may be possible to mold the predictors to fit the assumptions.

Finally, we have learned a fair bit from this exercise and as a consequence feel that graphical regression may have useful role in some practical problems.

References

Becker, R. A., Cleveland, W. S (1987). Brushing Scatterplots. *Technometrics*, **29**, 127-142.

Becker, R. A., Cleveland, W. S. and Wilks, A. R. (1987). Dynamic graphics for data analysis (with discussion). *Statistical Science*, **2**, 355-395.

Becker, R. A., Cleveland, W. S. and Weil (1988). The use of brushing and rotation for data analysis. In *Dynamic Graphics for Statistics,* W. S. Cleveland and M. E. McGill (eds), Belmont, CA: Wadsworth, p. 247-276.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis.* Boston: Duxbury Press.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cleveland, W. S. and McGill, M. E. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association*, **79**, 807-822.

Cleveland, W. S. (1987). Research in Statistical Graphics. *Journal of the American Statistical Association*, **82**, 419-423.

Cleveland, W. S. and McGill, M. E. (1988). *Dynamic Graphics for Statistics.* Belmont: Wadsworth.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Ser B*, 133-155.

Cook, R. D. (1987). Software review of MacSpin. *The American Statistician*, **41**, 233-236.

Cook, R. D. (1992). Graphical regression. In Dodge, Y. and Whittaker, J. W., *Computational Statistics, Vol 1*, Heidlberg: Physica-Verlag, p. 11-22.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, to appear.

Cook, R. D. and Nachtsheim, C. J. (1992). Re-weighting to achieve elliptically contoured covariates in regression. Working Paper 92-13, Department of Operations and Management Science, University of Minnesota, Minneapolis, MN 55455.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

Cook, R. D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphics (with discussion). *Technometrics*, **31**, 277-312.

Cook, R. D. and Weisberg, S. (1990). Three dimensional residual plots, in Berk, K. and Malone, L. eds., *Proceedings of the 21st Symposium on the Interface: Computing Science and Statistics.* Washington: American Statistical Association, 1990, 162-166.

Cox, D. R. (1978). Some remarks on the role in statistics of graphical methods. *Applied Statistics*, **27**, 4-9.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Ser B*, **30,** 248-275.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **41**, 1-31.

Eaton, M. L. (1986). A characterization of spherical distributions. *Multivariate Analysis*, **20**, 272-276.

Fisherkeller, M. A., Friedman, J. H. and Tukey, J. W. (1974). PRIM-9: An interactive multidimensional data display and analysis system. In *Dynamic Graphics for*

Statistics, (1988), W. S. Cleveland and M. E. McGill (eds), Belmont, CA: Wadsworth, p. 91–110.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, **74**, 503–515.

Hastie, T. J. and Tibshirani (1990). *Generalized Additive Models.* London: Chapman and Hall.

Huber, P. (1985). Projection pursuit (with discussion). *Annals of Statistics*, **13**, 435-525.

Huber, P. (1987). Experiences with three-dimensional scatterplots. *Journal of the American Statistical Association*, **82**, 448-453.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, **87**, 1025-1040.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics*, **17**, 1009-1052.

Tierney, L. (1990). *LISP-STAT.* New York: Wiley.

Young, F. W., Kent, D. P. and Kuhfeld, W. F. (1988). Dynamic graphics for exploring multivariate data. In *Dynamic Graphics for Statistics,* W. S. Cleveland and M. E. McGill (eds), Belmont, CA: Wadsworth, p. 391-424.

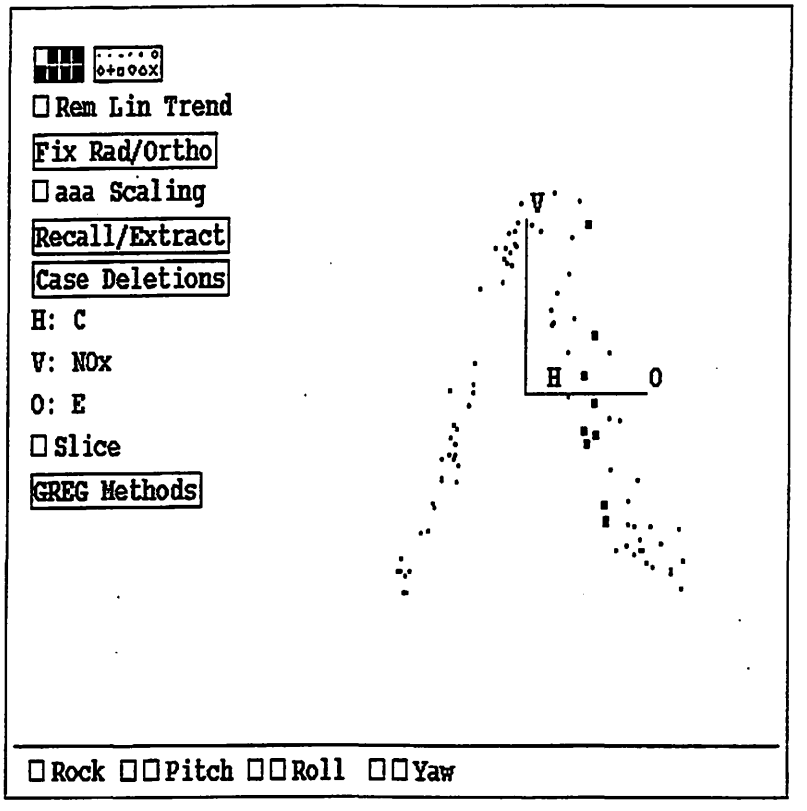

Figure 1: Ethanol data rotated to 0.01 C + 0.99 E with sliced points selected

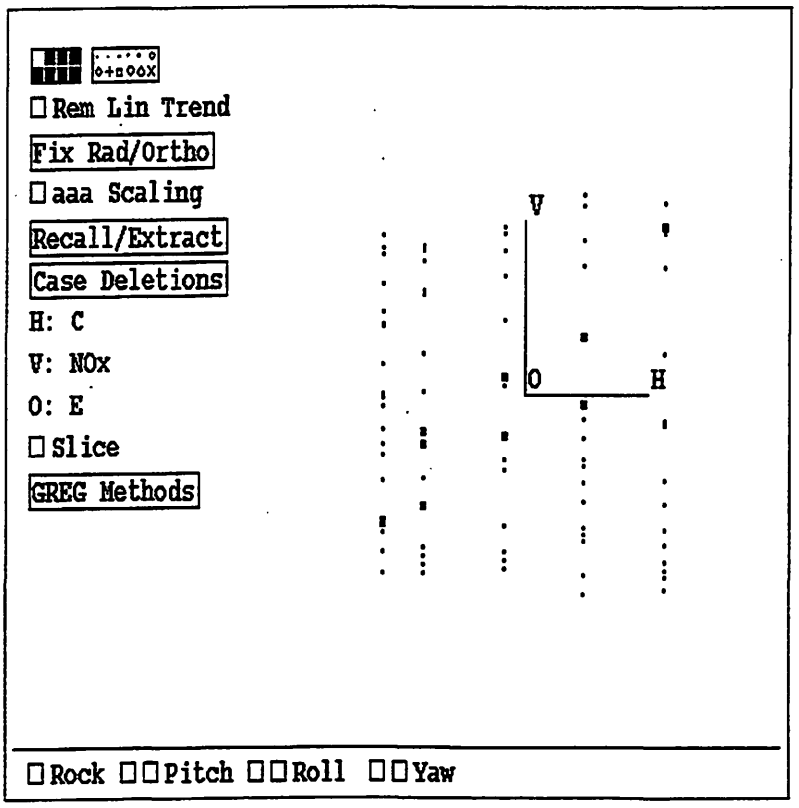

Figure 2: Figure 1 rotated to see the trend in a second direction. The same points are selected in Figure 1.

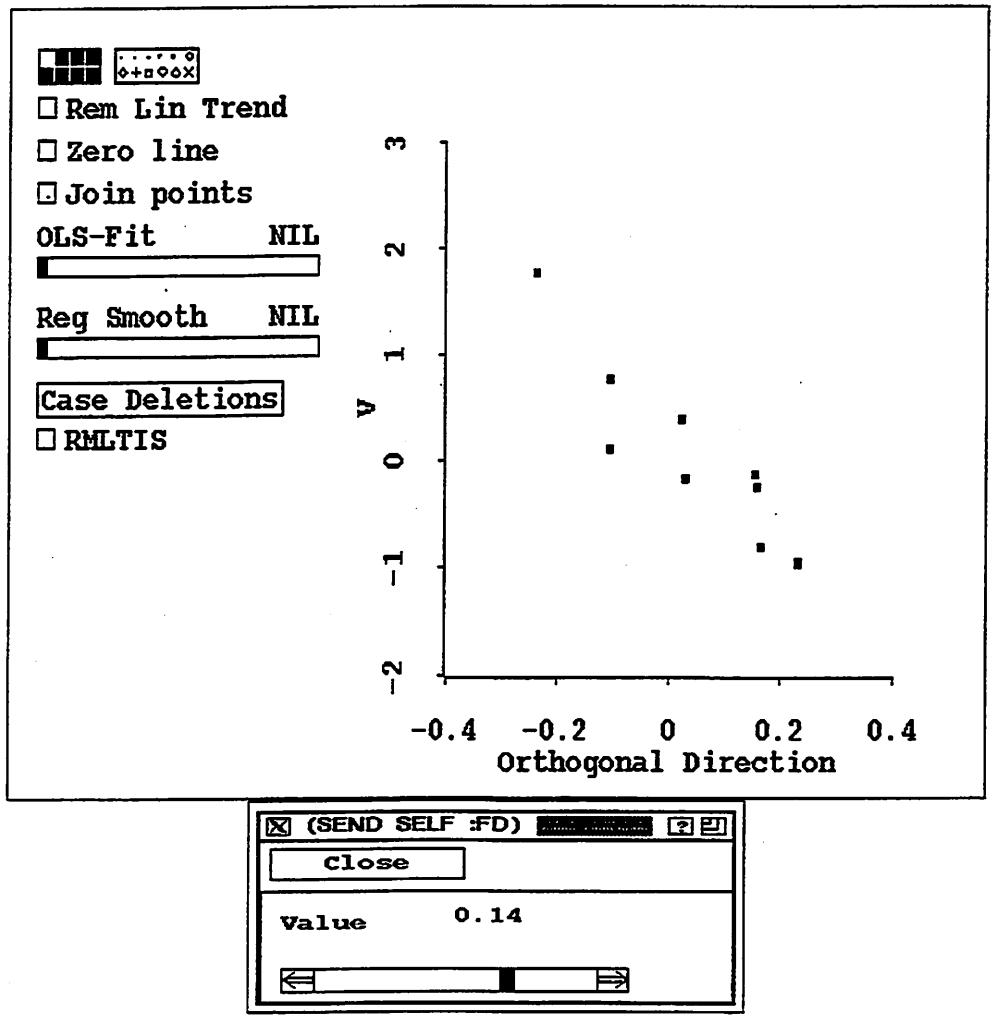

Figure 3: Selected data from Figures 1 and 2 plotted against the orthogonal direction.

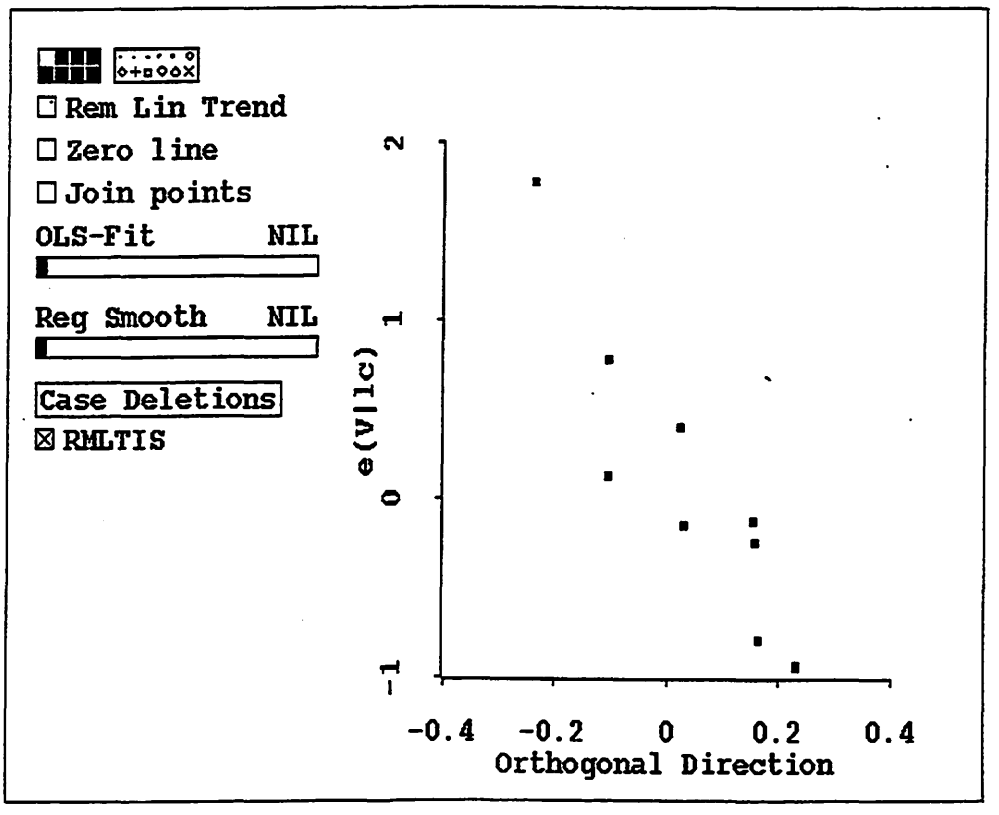

Figure 4: Selected data from Figures 1 and 2 plotted against the orthogonal direction with the linear trend removed.

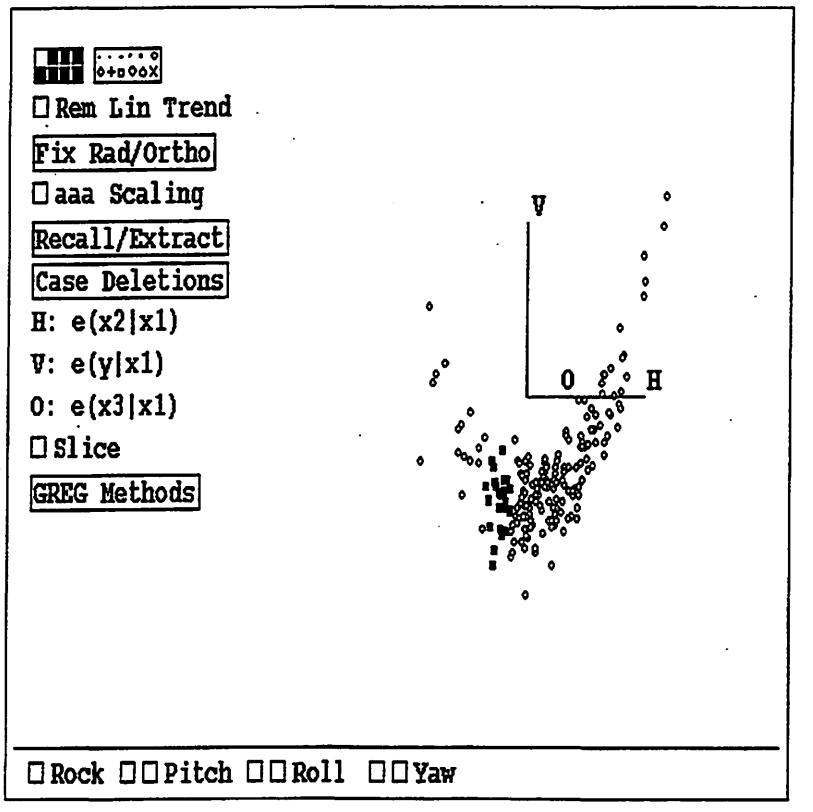

Figure 5: Added Variable Plot for Example 4.1: $\{e_{y|3}, (e_{1|3}, e_{2|3})\}$ rotated to the $.95e_{1|3} + .30e_{2|3}$ direction.

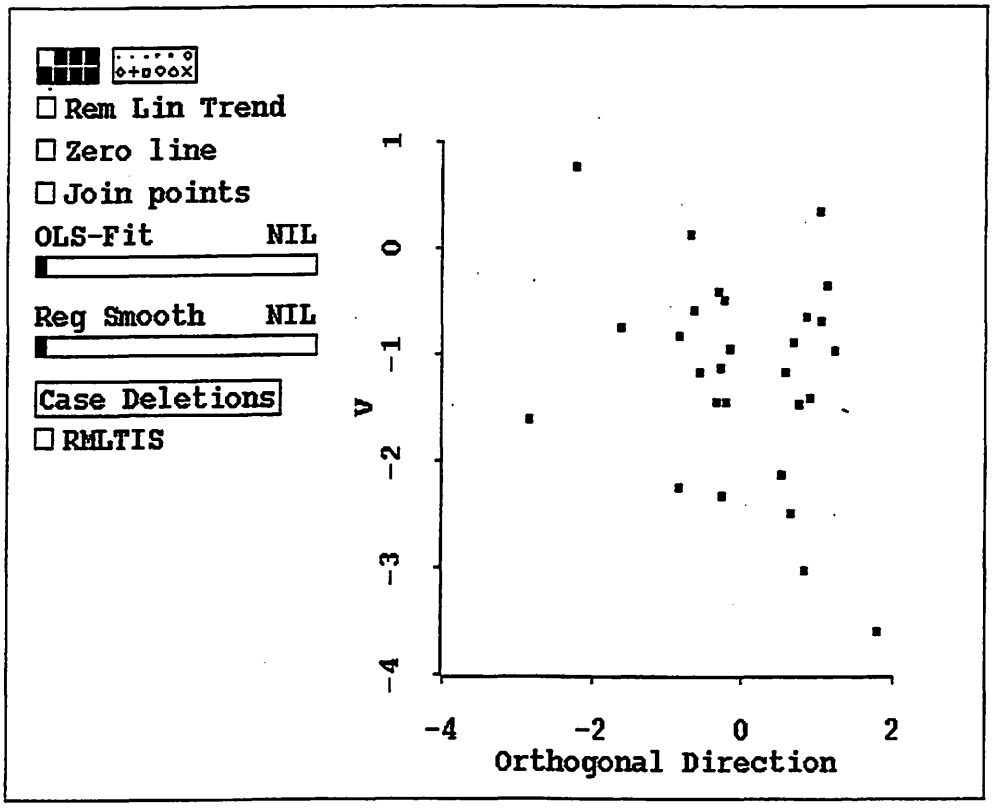

Figure 6: Selected points from Figure 5, plotted against an orthogonal direction.

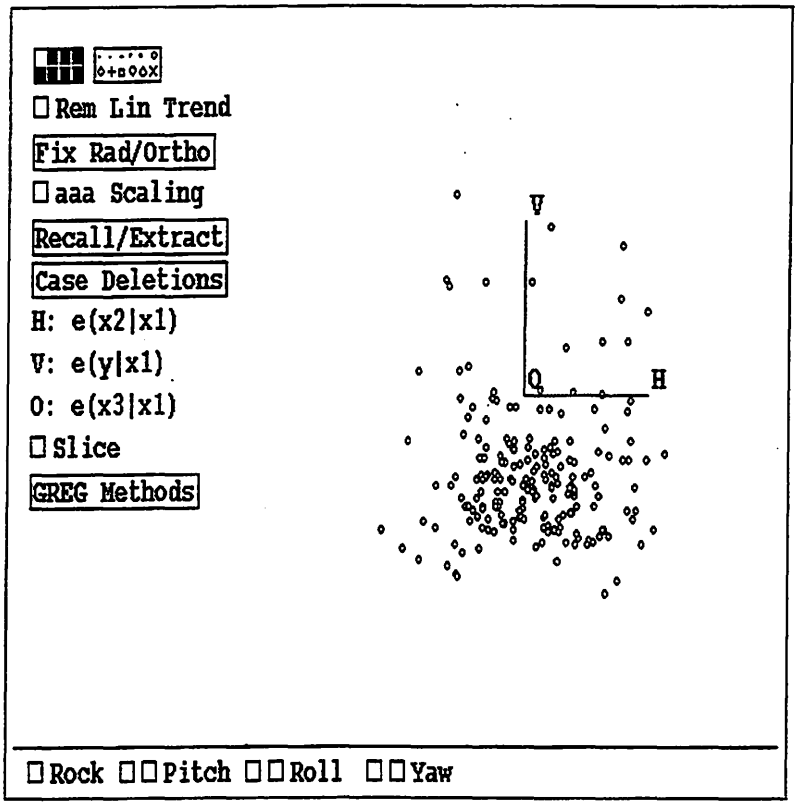

Figure 7: Added Variable Plot for Example 4.1: $\{e_{y|1}, (e_{2|1}, e_{3|1})\}$.

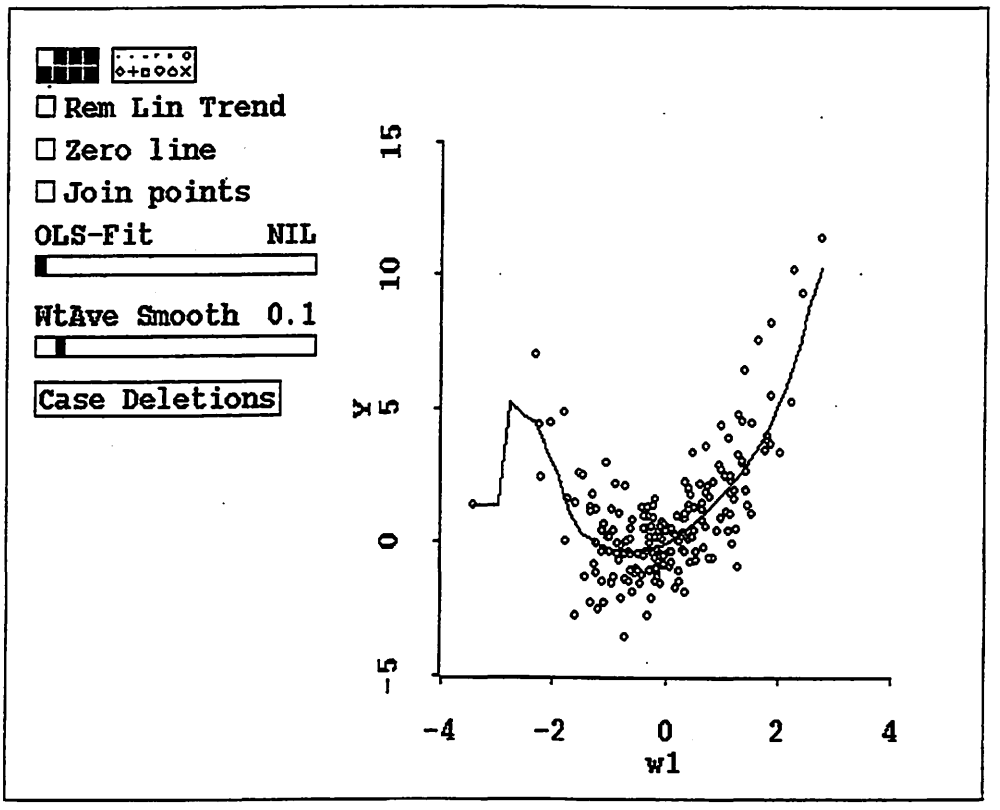

Figure 8: Plot of y versus w_1 with a triangle kernel smooth.

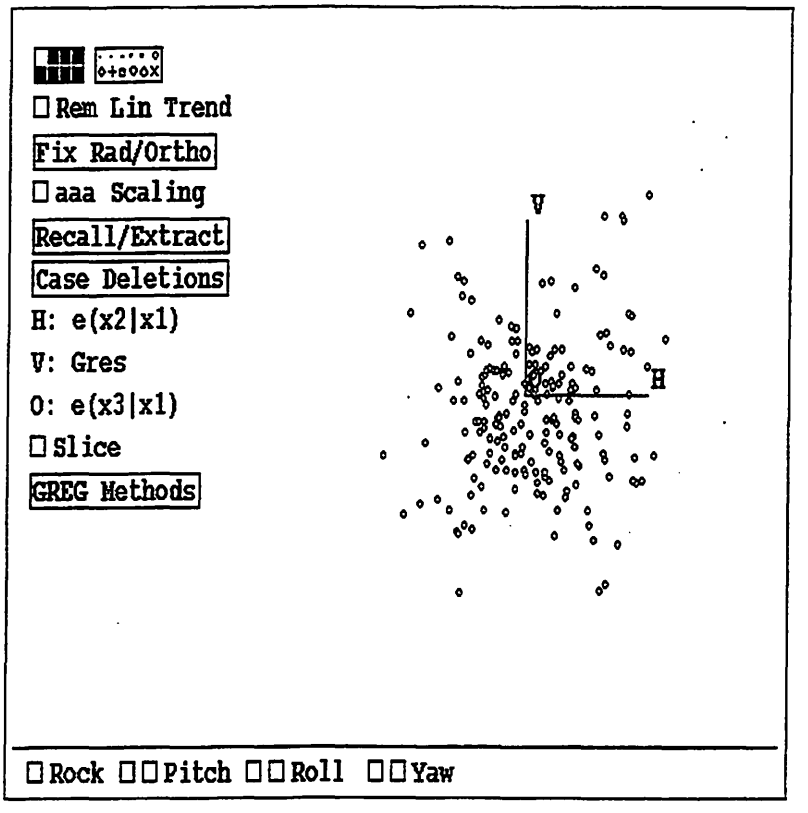

Figure 9: Added Variable Plot in Figure 7 with the vertical axis replaced by the residuals from the fit shown in Figure 8.

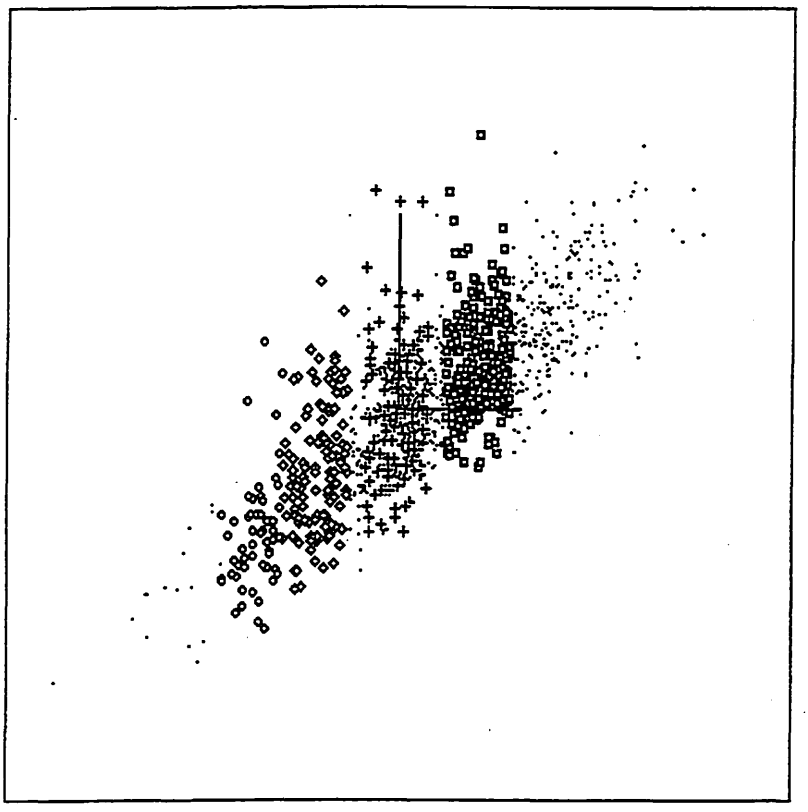

Figure 10: Added variable Plot for the ERA data with added predictors $\log w_3$ and $\log w_4$ rotated to the $.67 \log w_3 + .74 \log w_4$ direction. Point symobols correspond to the four slices in Figure 11.

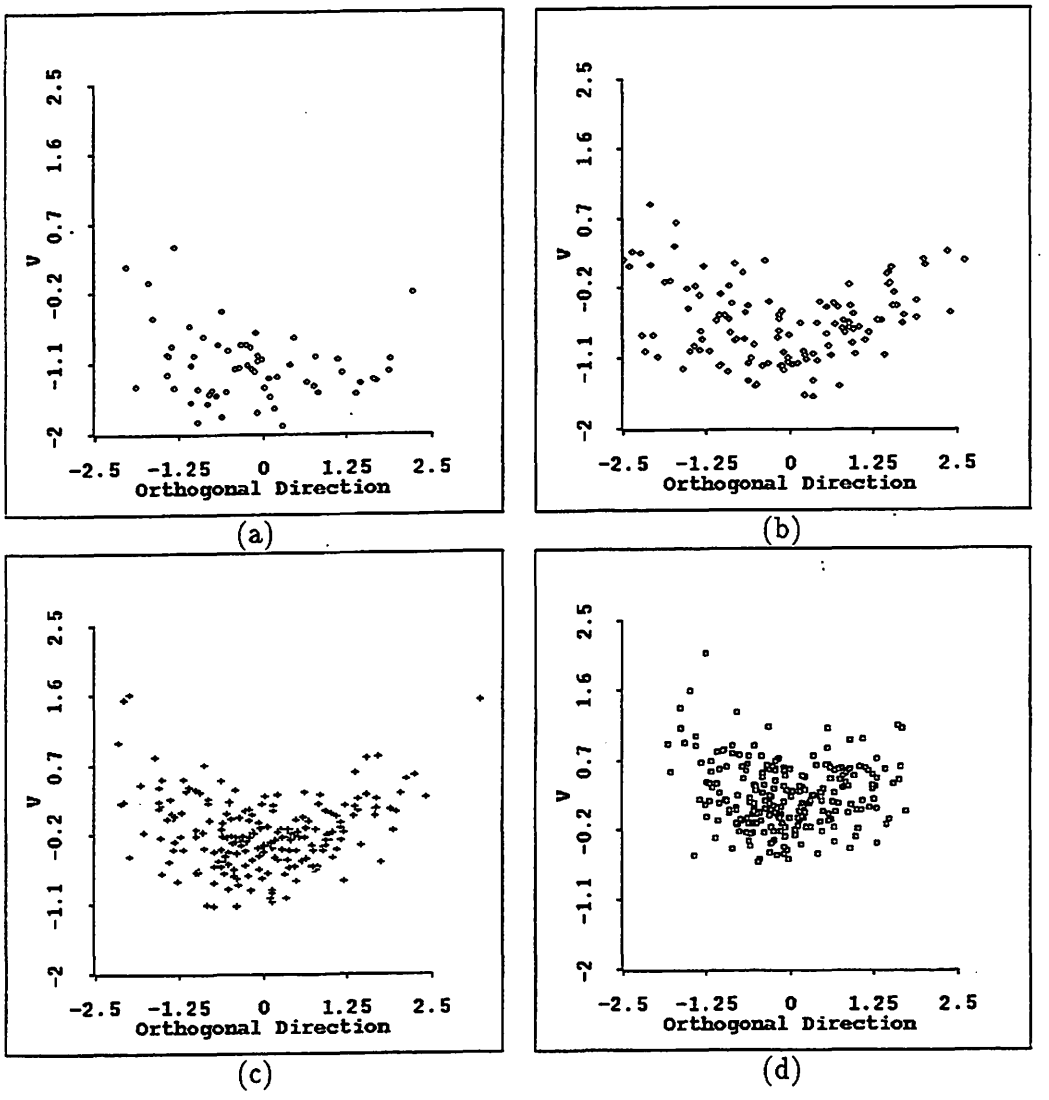

Figure 11: Four Slices of the data in Figure 10.

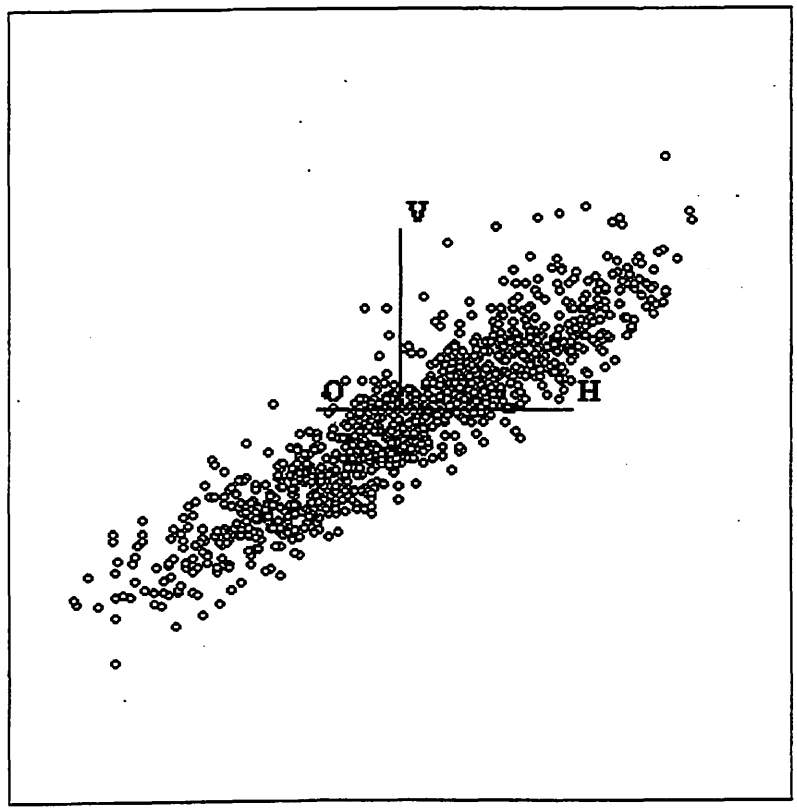

Figure 12: Added Variable Plot with added predictors $\log w_1$ and w_2. rotated to the $.01 \log w_1 - .99 w_2$ direction